ФОРМАЛИЗАЦИЯ OLAP-КУБОВ И РЕЛЯЦИОННЫХ СВЯЗЕЙ В БАЗАХ ДАННЫХ С ПОМОЩЬЮ АЛГЕБРЫ МНОЖЕСТВ

Трофимов С.П.[1], Пономарева О.А.[2]
[1] Уральский институт экономики, управления и права,
ул. Луначарского, 194, Екатеринбург, Свердловская обл., 620026, Россия
e-mail: tsp61@mail.ru
Уральский федеральный университет имени первого Президента России Б.Н. Ельцина
ул. Мира, 19, Екатеринбург, Свердловская обл.,620002, Россия,
e-mail: ponomareva1704@rambler.ru

Аннотация — Предлагается математическая формализация OLAP-кубов, в которой размерности куба и их декартовые произведения снабжаются алгеброй подмножеств и мерой. Эти множества участвуют в запросах. Операции проекции и сечения согласуются с алгебрами размерностей. Реляционные связи между размерностями реализуются с помощью индексных отображений на произведениях этих размерностей. Вводятся три типа индексных отображений на размерности, соответствующие трем типам реляционных связей. Показано, что реляционная база данных является OLAP-кубом с соответствующими индексными отображениями и может быть записана одной формулой. В качестве примера рассматривается аналитическая обработка анкет.

FORMALIZATION OF OLAP-CUBES AND RELATIONAL DATABASES WITH THE ALGEBRA OF SETS

Trofimov S.P.[1], Ponomareva O.A.[2]
[1] Ural Institute of Economics, Management and Law,
Lunacharskogo St, 194, Yekaterinburg, Sverdlovsk region, 620026, Russian Federation
e-mail: tsp61@mail.ru
[2] Ural Federal University named after the first President of Russia B.N.Yeltsin
Mira St, 19, Yekaterinburg, Sverdlovsk region, 620002, Russian Federation
ph.: (343) 375-44-68, e-mail: o.g.trofimova@mail.ru

Abstract — Mathematical formalization of OLAP-cubes is developed. Dimensions of cube and Cartesian products are supplied with an algebra of sets and measure. These sets are involved in queries. Operations projection and cross section are consistent with the algebra of dimension. Relational connections between dimensions realized with the help of index maps on the works of these dimensions. We introduce three types of index maps to the dimension corresponding to the three types of relational ties. It is shown that the relational database is an OLAP-cube with the corresponding index maps, and can be written by one formula. As an example the analytical processing of the questionnaires

I. Актуальность работы

Основными способами структурирования данных являются реляционные базы данных и иерархические базы данных.

Реляционные базы данных предполагают разбиение исходной информации и размещение в отдельных таблицах. Некоторые таблицы представляют собой сущности, а поля таблицы – атрибуты сущностей. Вспомогательные таблицы реализуют связь «многие-ко-многим». Реляционные базы удовлетворяют нормальным формам и позволяют максимально компактно хранить информацию [1]. Однако при этом возникают трудности при формировании и выполнении запросов.

Примером иерархических баз данных являются базы, построенные с использованием xml-формата. В данном случае необходимо создать синтаксический набор тегов, атрибутов и других элементов. Запросы реализуются с помощью технологии XPath.

Переход к OLAP-кубам [2] предоставляет третий способ структурирования данных, при котором сущности представляют собой подпространства с осями атрибутов. Кубы являются разреженными структурами, для которых запросы реализуются с помощью операций проекции и сечения с применением агрегатных функций. OLAP-кубы имеют прозрачную организацию данных, позволяют достаточно просто формировать и выполнять запросы.

Технология OLAP-кубов постоянно развивается и пополняется новыми алгоритмами. В статье [3] исследуются способы навигации и поиска информации в больших объемах данных. С этой целью рассматриваются многозначные триадические контексты, представляющие собой тройки, состоящие из объекта, атрибутов объекта и множеств значений атрибутов. Показано, что многозначные контексты можно представить, как размерность OLAP-куба. В работе [4] OLAP-кубы используются для представления триадических контекстов, в которых множества объектов, атрибутов и значений атрибутов представляют собой произвольные множества и являются взаимозаменяемыми.

Мы предлагаем алгебраический поход к определению OLAP-кубов, который представляет базовую размерность куба как тройку, состоящую из множества, снабженного алгеброй и мерой. Более сложные размерности, содержащие множество атрибутов, задаются произведениями базовых размерностей. Для индексации элементов размерности мы вводим индексные отображения, которые позволяют однотипно реализовать три основные связи между сущностями: «один-к-одному», «один-ко-многим» и «многие-ко-многим».

II. Основные проблемы и решения

Определение 1. Базовой размерностью будем называть произвольное множество A, с которым связана алгебра **A**, состоящая из некоторых подмножеств множества A, и мера μ: **A** → R. Таким образом, базовая размерность задается тройкой (A, **A**, μ). Множество A будем называть носителем размерности.

По определению алгебры множеств объединение конечного числа и пересечение произвольного числа элементов из A является также элементом этой алгебры.

Пример 1. A – конечное множество фирм, **A** – все подмножества из A, мера μ подмножества совпадает с мощностью этого подмножества.

Пример 2. Множество A представляет собой цены, то есть совпадает с множеством неотрицательных чисел, алгебра **A** – конечные интервалы, μ - длина интервала.

Для базы данных базовая размерность представляет собой сущность, для которой множество A является единственным атрибутом, элементы алгебры **A** используются в запросах, мера μ - агрегатная функция.

Рассмотрим две базовые размерности R_1 = (A, **A**, μ_A) и R_2 = (B, **B**, μ_B). На декартовом произведении C=A×B можно задать несколько различных размерностей, если в C выбрать различные алгебры подмножеств и меры на них.

Пример 3. Элементы алгебры **C** можно составить из декартовых произведений A_1×B_1 элементов A_1 из **A** и B_1 из **B** и их конечных объединений. Мера на **C** является произведением мер на **A** и **B**: $\mu_C(A_1 \times B_1) = \mu_A(A_1) \cdot \mu_B(B_1)$.

Пример 4. Плоскость Oxy представляет собой множество $C = R^1 \times R^1$. В качестве порождающих элементов алгебры на **C** можно задать прямоугольники. Мерой прямоугольника является его площадь. Проекции и сечения прямоугольников совпадают с алгеброй интервалов, заданной на прямых.

В качестве порождающих элементов алгебры можно выбрать параллелограммы. В этом случае в запросах могут участвовать взвешенные линейные комбинации двух атрибутов. Такой пример является очевидным для нашего подхода. В практике OLAP-кубов он не встречается, хотя его практическая значимость очевидна. Например, запрос на автомобили, у которых цена на автомобиль и сумма цены автомобиля и цены его обслуживания находятся в заданных диапазонах, порождает параллелограмм в двумерной базовой координате сущности «автомобиль» с двумя атрибутами «цена автомобиля» и «цена обслуживания».

Определение 2. Рассмотрим конечное или счетное подмножество S из носителя некоторой базовой размерности. Будем считать, что точки множества S проиндексированы, если задано взаимно однозначное отображение $i = i(t)$ элементов t из S в множество целых чисел. Таким образом, каждому t из S ставится в соответствие некоторое целое число так, что из $i(t_1) = i(t_2)$ вытекает $t_1 = t_2$.

По известному индексу $i(t)$ можно восстановить элемент t носителя размерности $t = i^{-1}(i(t))$.

Пример 5. Элементы одномерного массива индексируются своим индексом, открытые файлы индексируются файловыми дескрипторами, элементы двумерного массива индексируются смещением относительно начала массива.

Допустим, имеются две проиндексированные базовые размерности R_1 = (A, **A**, μ_A) и R_2 = (B, **B**, μ_B) с индексами i_A и i_B. Создадим новую размерность R_3 = (C, **C**, μ_C) с помощью произведения базовых размерностей R_1 и R_2. В качестве носителя для R_3 возьмем C = A×B. Зададим произвольным образом для размерности R_3 некоторую алгебру подмножеств **C** и меру μ_C. Построим индексацию $i(c)$ точек c=(a, b) из C несколькими способами:

I1. $i(c) = i(i_A(a), i_B(b))$. Элемент c=(a, b) индексируется одним целым числом, которое образуется с помощью некоторого алгоритма из двух целых чисел $i_A(a)$ и $i_B(b)$.

I2. $i(c) = i(a, b)$ Индекс формируется для пары исходных объектов (a, b). При этом индексы размерностей $R1$ и $R2$ не используются.

I3. $i(c) = i(i_A(a), b)$ или $i(c) = i(a, i_B(b))$. При индексации элемента с используется индекс одной из размерностей и элемент другой размерности.

Способ индексации I1 будем обозначать A*B. Он позволяет связать размерности R_1 и R_2 отношением «многие-ко-многим» (Рис.1)

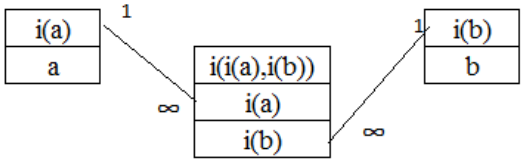

Рис. 1. Произведение размерностей R_1 и R_2 с отношением многие-ко-многим

Вариант I2 будем обозначать AoB. Он позволяет связать размерности R_1 и R_2 отношением «один-к-одному» (Рис.2)

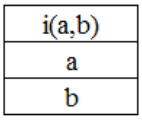

Рис. 2. Произведение размерностей R_1 и R_2 с отношением один-к-одному

Третий вариант I3 будем обозначать A→B. Он позволяет связать размерности R_1 и R_2 отношением «один-ко-многим» (Рис.3).

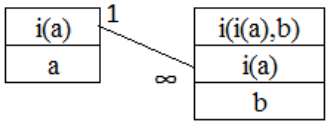

Рис. 3. Произведение размерностей R_1 и R_2 с отношением один-ко-многим

Во всех трех вариантах создания новой размерности R_3 носитель C, на котором определена эта размерность, является одним и тем же декартовым произведением носителей A и B. Таким образом, размерность R_3 характеризуется не только несущим множеством, алгеброй подмножеств и мерой, но также способом индексации. Причем

способ индексации не зависит от других свойств размерности.

Определение 3. Пусть R_1, … ,R_n – некоторые неиндексированные базовые размерности. Рассмотрим произвольное отображение W декартова произведения $A = A_1 \times \ldots \times A_n$ носителей этих размерностей в бинарное множество {0, 1}. OLAP–кубом называется множество $W^{-1}(1)$, являющееся подмножеством A.

Таким образом, OLAP–куб является подмножеством декартова произведения $A_1 \times \ldots \times A_n$. Допустим на размерностях $R_1,\ldots,R_n$ заданы индексы. Для индексации $i(a)$ произвольной точки $a = (a_1,\ldots,a_n)$ этого куба можно выбрать различные способы. Допустим индекс $i(a)$ вычисляется на основе индексов отдельных размерностей, то есть $i(a) = i(i(a_1),\ldots,i(a_n))$. В этом случае OLAP-куб является произведением индексированных размерностей, построенном по способу I1, то есть задается выражением $A_1 * \ldots * A_n$.

Отметим, что произведение всех базовых размерностей является тоже размерностью.

III. Связь между OLAP-кубами и реляционными базами данных

Рассмотрим упорядоченную последовательность из базовых размерностей R_1, …, R_n . Расставим между ними знаки операций {*, o, →} произвольным образом. Приоритеты операций будем указывать круглыми скобками. По умолчанию приоритет слева на право.

Каждая из трех операций создает некоторые таблицы, а все выражение задает некоторую схему базы данных. Каждая схема базы данных задает свою организацию базы данных в виде физических таблиц и связей между ними. В свою очередь, каждая база данных может быть представлена в виде одной плоской таблицы или виде OLAP- куба в декартовом произведении $R_1 \times \ldots \times R_n$.

Определение 4. Две базы данных эквиваленты если они представимы в виде одного и того же OLAP куба.

Пример 6. Построим выражение $(R_1 * R_2) \rightarrow (R_3 o R_4)$ В соответствии с вышеописанным алгоритмом оно задает следующую схему базы данных (рис.4).

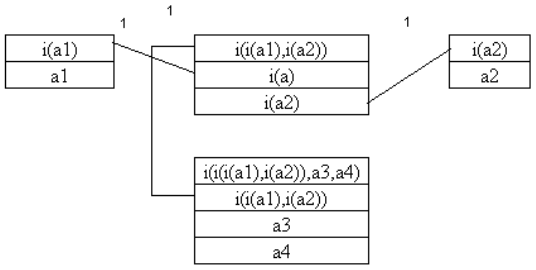

Рис.4 Схема базы данных выражения $(R_1 * R_2) \rightarrow (R_3 o R_4)$

Несмотря на сложный вид, эту схему нельзя упростить. Из схемы видно, что от индекса $i(a_3,a_4)$ мы отказались.

Пример 7. Схема базы данных для выражения $R_1 \rightarrow (R_2 \rightarrow R_3) \rightarrow R_4$ изображена на рисунках 5 и 6.

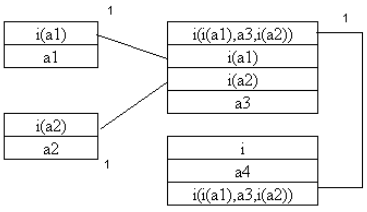

Рис. 5 Схема базы данных выражения $R_1 \rightarrow (R_2 \rightarrow R_3) \rightarrow R_4$

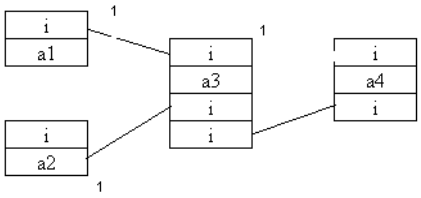

Рис 6 Упрощенная схема базы данных выражения $R_1 \rightarrow (R_2 \rightarrow R_3) \rightarrow R_4$

Пример 8 Схема базы данных для выражения $((R_1 \text{ o } R_2) \text{ o } R_3) * R_4$ изображена на рис. 7.

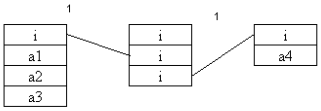

Рис 7 Упрощенная схема базы данных выражения $((R_1 \text{ o } R_2) \text{ o } R_3) * R_4$

IV. Применение подхода при обработке результатов анкетирования

Применение OLAP-кубов актуально при обработке больших массивов данных. Применим изложенный подход к обработке результатов анкетирования.

В УрФУ в течение 2011 и 2012гг. проводится анкетирование по качеству обслуживания в комбинате питания университета. Приведем некоторые вопросы анкеты.

1. Как часто вы пользуетесь услугами столовых? Варианты ответов: Каждый день; 3 раза в неделю; Реже 2 раз в неделю; Не пользуюсь услугами столовых УрФУ.

2. Сколько времени Вы тратите на обед во время большой перемены с учетом ожидания в очереди

	<10 мин.	10-20 минут	20-30 минут	>30 мин.
Столовая теплофака				
……				
Буфеты				

9. Ваш возраст: до 20 лет; 20-24 года; 25-29 лет; 30-34 года; 35-39 лет;40-44 года; 45-49 лет; 50 –54 года; 55-59 лет; 60 и старше.

10. Сообщите, пожалуйста, какова средняя стоимость вашего обеда. от 30 до 50 руб.; от 50 до 70 руб.; от70 до 90 руб.; от 90 до 110 руб.; 110 и более руб.

В анкете каждый вопрос представляет собой название отдельной размерности, а варианты ответов несущие множества этой размерности. Назовем эти размерности B_1, …B_{10}.

Пример организации схемы базы данных для обработки анкетных данных

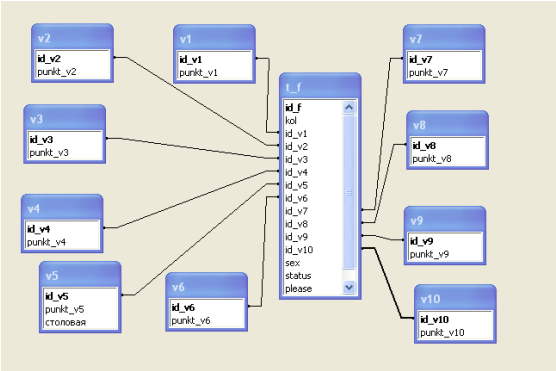

Рис. 8. Схема базы данных анкетирования

Этой схеме базы данных соответствует запись $R_1*R_2*\ldots*R_{10}$, при условии, что вопросы имеют однозначный ответ.

Возможны вопросы с множеством вариантов ответов, например, вопрос №2. В данном случае размерность R_2 образуется из произведения двух размерностей: S_1 – место нахождения столовой, S_2 – продолжительность обеда. Для индексации используется способ I3: $R_2=S_1\rightarrow S_2$. В этом случае схема базы данных задается выражением

$R_1* (S_1->S_2) * R_3*\ldots*R_{10}$.

V. Заключение

В статье предложено математическое описание размерностей для построения OLAP-кубов. Применение различных способов индексации для новых размерностей позволяет конструировать различные схемы реляционных баз данных.

VI. Литература

[1] Гарсиа-Молина Г. Системы баз данных/ Г.Гарсиа-Молина, Дж.Ульман, Дж. Уидом. – М.: Вильямс, 2004. – 1088 с.

[2] Бергер А. OLAP и многомерный анализ данных / А. Бергер, И.Горбач. СПб.: БХВ-Петербург, 2007. – 905с.

[3] Ferré, S., Allard, P., Ridoux, O.: Cubes of Concepts: Multi-dimensional Exploration of Multi-valued Contexts. ICFCA 2012: 112-127

[4] Stumme, G: A Finite State Model for On-Line Analytical Processing in Triadic Contexts. In: Formal Concept Analysis. Lecture Notes in Computer Science Volume 3403, 2005, pp 315-328